# 文書のチャンクに基づく知識グラフを活用した RAG

江上周作 [1]　福田賢一郎 [1]
[1] 国立研究開発法人産業技術総合研究所 人工知能研究センター
{s-egami,ken.fukuda}@aist.go.jp

## 概要

大規模言語モデルで外部知識を活用する上で Retrieval-Augmented Generation (RAG) が注目を集めており，RAG の各モジュールごとに拡張手法が開発されている．検索モジュールでは埋め込みベクトルに基づく類似度検索手法の開発が主流となっているが，ユーザの質問に答える上で必要な情報は必ずしも質問に類似しない．本研究では，外部文書のチャンクのメタデータやチャンク間の意味的関係を抽出することで知識グラフを構築し，ベクトル類似度検索と組み合わせたハイブリッドな RAG 手法を提案する．実験の結果，ベースラインと比較して回答の忠実性とコンテキストの関連性の向上を確認した．

## 1　はじめに

大規模言語モデル（Large Language Model: LLM）を活用したシステムは質問応答を始め様々なタスクにおいて最先端の性能を凌駕し，その汎用性の高さや開発基盤の普及も相まって，LLM を活用したサービスの社会実装が急速に進んでいる．しかしながら，LLM の学習データに存在しない知識に関する質問応答ではハルシネーションの課題があり，その解決策として外部知識の活用の研究が進んでいる．中でも，Retrieval-Augmented Generation (RAG) [1, 2] は導入コスト，知識の更新，情報源の提示などの面で優れており注目を集めている．RAG の定義は技術の進展とともに拡大しているが，ここでは Gao ら [2] の定義を参照し，質問応答やテキスト生成の際にまず膨大な文書コーパスから関連する情報を検索し，その後取得した情報を利用して回答やテキストを生成し，それにより LLM の予測の質を高める手法とする．

図 1 に一般的な RAG 手法の全体図を示す．これにより LLM に最終的に入力されるプロンプトに追加される文脈情報は質問に類似する文章となる．しかしながら，質問に応答する上で必要な知識は，必ずしも質問に類似するチャンクに含まれるとは限らない．ユーザの要求の種類は単に主語や目的語を問う質問だけでなく，原因や結果を問う質問も想定され，これらの質問に回答する上で必要なチャンクは，質問の類似度の高いチャンクは分割されている可能性がある．

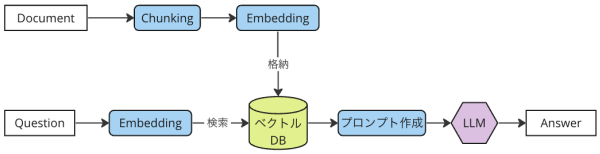


**図 1**　一般的な RAG 手法の全体図

本研究では，文書のチャンクを構造化することにより，LLM の予測精度向上に寄与する文脈情報を取得できるという仮説を立て，チャンク間の意味的関係に基づいて知識グラフを構築する．この知識グラフを用いた検索と埋め込みベクトルの類似度検索を組み合わたハイブリッド検索の RAG 手法を提案する．知識グラフは各チャンクの 5W1H 情報とチャンク間の談話関係 (Discourse Relation) を含む．質問文の埋め込みベクトルを用いた類似度検索により取得したチャンクから，知識グラフ検索で談話関係を辿ることで関連するチャンクを取得する．また，質問文から抽出した 5W1H 情報をもとにグラフ検索によりチャンクを取得する．取得したすべてのチャンクをリランクしてプロンプトを作成し，LLM へ入力することで文脈に沿った回答を得る．実験の結果，ベースラインと比較して回答の忠実性とコンテキストの関連性の向上を確認した．本稿では現時点での限界と今後の展開について述べる．

## 2　関連研究

RAG における文書の検索前，埋め込み時，検索後の各プロセスごとに様々な手法が提案されている [2]．検索前の処理としてテキストから不必要な情報の除去，メタデータの導入，仮説質問の導入，キーワード検索やセマンティック検索との組

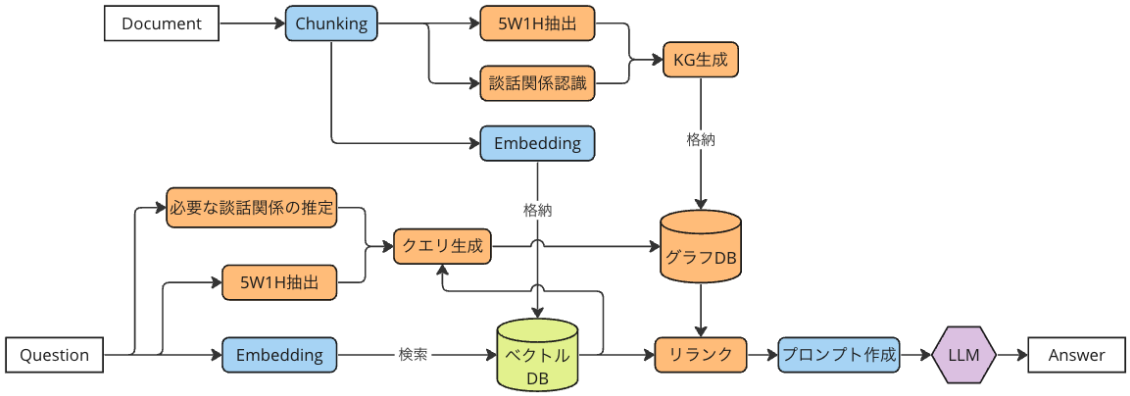


**図 2**　提案手法の全体図

み合わせなどの手法が存在する [2]．RAG における質問応答，会話，ツール検索などの多様なタスクに統一的に対応可能な埋め込みモデルとして，LLM-Embedder [3] が提案されている．検索後の処理では，コンテキストの最初か最後に関連情報が出現すると性能が高くなることが明らかになっており [4]，オープンソースライブラリの HeyStack 等でリランクのモジュールとして実装されている [5]．その他，各プロセスで様々な手法が提案されているが，いずれもユーザの質問に答える上で必要な情報は質問に類似するという前提であり，これに対して本研究ではチャンク間の意味的関係や構造情報を加味した検索を導入する．

# 3　提案手法

本研究の提案手法の全体図を図 2 に示す．橙色で着色した箇所は図 1 から新たに追加された機能である．提案手法による RAG の拡張は主に次の 3 点であり，本章で説明する．

1. 文書のチャンクに基づく知識グラフの構築
2. ベクトル DB とグラフ DB のハイブリッド検索
3. 検索結果のリランク

## 3.1　文書のチャンクに基づく知識グラフの構築

RAG では外部文書をデータベースに取り込む際，一定のトークン長からなるテキストの塊（チャンク）に分割し，チャンクごとに埋め込みベクトルを生成する．本研究では，このチャンクのテキストから知識グラフを構築する．

### 3.1.1　知識グラフのスキーマ

文から主語，述語，目的語の三つ組み（トリプル）を抽出して知識グラフを構築すると，二項関係

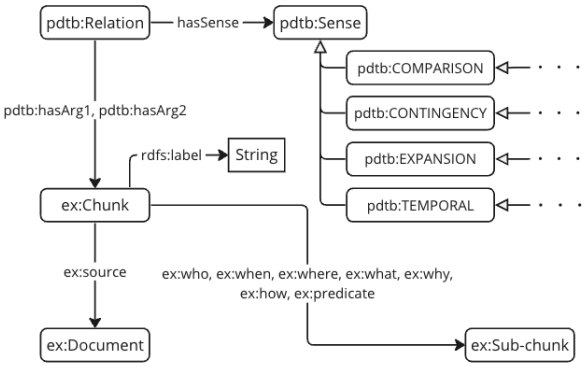


**図 3**　知識グラフのスキーマの一部

以上の意味を含む文の場合に，時間や場所などの重要な情報が欠落する．本研究ではナレッジグラフ推論チャレンジ[1) ] [6, 7, 8] で提供されている知識グラフのスキーマを参考にし，チャンクごとに “Who 誰が”，“Why(いつ)”，“Where(どこで)”，“What(何を)”，“Why(なぜ)”，“How(どのように)” の 5W1H 情報と，“Predicate(述語)” を記述する．

また，チャンク間の意味的な関係を記述するために Discourse Relation (談話関係) を使用する．談話関係の語彙体系として OLiA Annotation Model for PTDB relations (以降 PDTB オントロジー) [9] を使用する．これは，談話関係のアノテーションデータセット Penn Discourse Treebank [10] で使用されている談話関係語彙のオントロジーであり，談話関係クラスの階層関係や引数構造が定義されている．

図 3 に構築する知識グラフのスキーマの一部を示す．知識グラフは Resource Description Framework (RDF) 形式のデータとして構築する．角丸矩形はクラス，矩形はリテラルを意味する．矢印はプロパティ，白三角矢印は上位クラスを意味する．

1)　https://challenge.knowledge-graph.jp/

#### 3.1.2 5W1H 情報の抽出

5W1H 情報の抽出はイベント抽出の研究としてニュース記事，ウェブページ，物語などを対象に様々な手法が提案されている [11, 12, 13]．しかし，普遍的に利用可能な実装が公開されていない問題があり，オープンソースの 5W1H 抽出器の Giveme5W1H [14] が開発された．本研究ではまず Giveme5W1H を適用して 5W1H 情報の抽出を施行したが，抽出される単語と文節の精度や語彙統制に大きな課題があることが判明した．これは，入力されるチャンクが必ずしも一文の体裁を成していないことが原因の一つである．

そこで，本研究では LLM を用いて Zero-shot Learning で 5W1H 情報を抽出した．付録 A の Listing 3 に 5W1H を抽出するためのプロンプトを示す．5W1H の内，該当する情報がない場合は答えないように指示し，また，語彙の統制のために可能な限り短い単語を使用するよう指示する．

#### 3.1.3 談話関係認識

暗黙的な談話関係認識の包括的な調査 [15] によると，事前学習済みの言語モデルに基づく手法が最先端である事がわかる．しかし，これらの公開済みの実装は，入手に契約を要するデータセットでの学習が必要であり，再利用の障壁が高い．そこで，3.1.2 節と同様に LLM を用いて Zero-shot Learning で談話関係を認識する．付録 A の Listing 4 に談話関係を認識するためのプロンプトを示す．プロンプトには PDTB オントロジーから抽出した談話関係クラスの内，サブクラスを持たない最下層のクラス名をリストとして組み込む．このリストの中から談話関係を選択することを指示し，可能な限りの語彙統制を図る．

### 3.2 ベクトル DB とグラフ DB のハイブリッド検索

#### 3.2.1 埋め込みベクトルの類似度検索

チャンクの埋め込みベクトルの生成に LLM-Embedder [3] を使用する．質問文についても同様に LLM-Embedder を使用して埋め込みベクトルを生成する．質問文の埋め込みベクトルとベクトル DB に格納されたチャンクの埋め込みベクトルの cosine 類似度を計算し，上位 $k$ 個のチャンクを質問応答に必要なコンテキストとして取得する．提案手法では $k = 1$ とする．

#### 3.2.2 知識グラフを用いた談話関係の検索

質問文を対象とした談話関係認識とは異なり，質問の意図や背景の推定に近いタスクである．そこで，LLM を用いた Zero-shot Chain-of-Thought プロンプティング [16] により，グラフ検索クエリに使用する談話関係を推定する．付録 A の Listing 5 にプロンプトのテンプレートを示す．プロンプトには PDTB オントロジーから抽出した談話関係名のリストが組み込まれる．

次に，3.2.1 節で取得したチャンクの ID と推定した談話関係を使用して，知識グラフの検索クエリを作成する．知識グラフは RDF 形式であるため，付録 A の Listing 1 に示す SPARQL クエリテンプレートを使用する．このクエリは，「特定のチャンクとの関係があり，その関係の意味が特定の談話関係クラス，またはその全ての子孫クラス」に該当するチャンクを取得するものである．

#### 3.2.3 5W1H 情報を用いた知識グラフ検索

この検索では，些末な情報を捨象してチャンクを表現する 5W1H のみを利用することで，類似度検索では取得できない可能性のあるチャンクを補完的に取得する．まず，質問文から抽出した 5W1H 情報を部分的に持つチャンクリソースを，一致するトリプル数の降順で取得する．質問文からの 5W1H の抽出方法は 3.1.2 節と同様である．抽出結果に応じて付録 A の Listing 2 の SPARQL クエリテンプレートを埋める．

### 3.3 検索結果のリランク

このモジュールでは，各検索モジュールで取得したチャンクについて，コンテキストとしてプロンプトに組み込む際の順序を決定する．本研究では，ベクトル類似度検索で取得したチャンクと談話関係のあるチャンクを連続して配置する．したがって，ベクトル類似度検索 (3.2.1 節) により取得したチャンクのリスト $L_v$，談話関係のグラフ検索 (3.2.2 節) により取得したチャンクのリスト $L_d$，5W1H のグラフ検索 (3.2.3 節) により取得したチャンクのリスト $L_w$ の順序は $L_v \prec L_d \prec L_w$ とする．

## 4 実験

ベースライン手法として図 1 の一般的な RAG を使用し，提案手法と性能を比較する．

### 4.1 実験設定

本実験では RAG のパイプラインを end-to-end で評価する ragas[2] [17] を使用して，回答の正確性と忠実性，コンテキストの関連性を評価する．ragas は人手で作成されたプロンプトに基づく自動評価フレームワークであり，人間による評価の代替として高性能な LLM を使用する．ragas は特に忠実性の評価において高性能であり，人間の予測と密接に一致することが示されている．本実験では ragas の LLM として GPT-4 を使用する．

データセットに Financial Opinion Mining and Question Answering (fiqa) dataset [18] を元に作成された 30 個の質問，Ground truth (GT)，文書のセット[3] を使用する．

文書のチャンクの最大トークン数は 50，オーバーラップは 0 とする．チャンク及び質問文の埋め込みモデルは LLM-Embedder を使用する．5W1H の抽出に使用する LLM は GPT-3.5 Turbo，談話関係認識および質問の回答に必要な談話関係の推定には GPT-4 Turbo[4] を使用する．最終的なプロンプトの入力先となる LLM は GTP-4 Turbo を使用する．

### 4.2 実験結果

実験結果を表 1 に示す．結果として，ベースラインと比較して提案手法では回答の忠実性とコンテキストの関連性が向上することが確認できた．回答が忠実であるということは，LLM で生成された回答の主張がコンテキストから推測できることを意味する．コンテキストの関連性は，取得したコンテキストのうち，質問に答える上で関連するコンテキストの割合を意味する．つまり，提案手法は回答に必要な情報をより無駄なく取得し，その結果に忠実に回答する手法と言える．したがって，回答のみでなくその根拠となる情報源を同時に提示する場合に有用であると考えられる．

一方で，回答の正確性はわずかに低下することがわかった．回答の正確性は，LLM の回答と GT との類似度と，回答と GT との F1-score の荷重平均を計算している．提案手法では，質問時に推定した談話関係を使用して知識グラフ検索を行う際，該当する関係が存在せずチャンクを取得できないケースが発生した．これにより，最終的なコンテキストの量が減少し，正確性の低下に影響する結果になったと考察する．各モジュールのアブレーションスタディや，言語モデルやリランク手法の違いによる詳細な分析については今後の課題とする．

**表 1** 実験結果

| | 正確性 | 忠実性 | コンテキストの関連性 |
|---|---|---|---|
| ベースライン | 0.469 | 0.199 | 0.365 |
| 提案手法 | 0.445 | 0.389 | 0.486 |

### 4.3 課題

知識グラフの構築において，5W1H 情報の抽出時にエンティティの統合に課題がある．本研究では，抽出結果の小文字化やレンマ化など，最低限の語彙の統制は行ったが，抽出結果は複数トークンからなるサブチャンクであることが多い．したがって，KG のエンティティの統一化が困難なケースが多く発生した．これにより，質問文から抽出した 5W1H 情報を利用したグラフ検索で十分に効力を発揮できていない．

また，LLM を使用した談話関係認識では結果に偏りがあることを観測した．全談話関係 3372 件のうち，1976 件は juxtaposition であり，一方で hypothesical のように 1 件も存在しない関係がある．データセットとして使用した fiqa は数百から数千文字の比較的短い文書であり，談話関係が十分に含まれていないことや，チャンク間に談話関係が無いことを正しく認識できていないことが原因として考えられる．これにより，質問から回答する上で必要な知識との談話関係を推定するプロセスにおける結果との整合性が取れず，グラフ検索で取得できるチャンク数が減少している．

## 5 おわりに

本研究では，チャンクから抽出した 5W1H 情報やチャンク間の談話関係を元に知識グラフを構築し，埋め込みベクトル類似度検索とグラフ検索を組み合わせた RAG 手法を提案した．これにより，回答の忠実性とコンテキストの関連性が向上することを確認した．今後，より詳細な実験を追加するとともに，明らかになった課題を解決する．

2) https://github.com/explodinggradients/ragas
3) https://huggingface.co/datasets/explodinggradients/fiqa/viewer/ragas_eval
4) https://openai.com/blog/new-models-and-developer-products-announced-at-devday

## 謝辞

本研究は JSPS 科研費 19H04168，22K18008，23H03688 の助成を受けたものです. 成果の一部は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務 JPNP20006 の結果得られたものです.

## 参考文献

[1] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 9459–9474. Curran Associates, Inc., 2020.

[2] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. Retrieval-Augmented Generation for Large Language Models: A Survey, December 2023. arXiv:2312.10997 [cs].

[3] Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. Retrieve Anything To Augment Large Language Models, October 2023.

[4] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts, November 2023. arXiv:2307.03172 [cs].

[5] Vladimir Blagojevic. Enhancing RAG Pipelines in Haystack: Introducing DiversityRanker and LostInTheMiddleRanker, August 2023.

[6] Takahiro Kawamura, Shusaku Egami, Koutarou Tamura, Yasunori Hokazono, Takanori Ugai, Yusuke Koyanagi, Fumihito Nishino, Seiji Okajima, Katsuhiko Murakami, Kunihiko Takamatsu, Aoi Sugiura, Shun Shiramatsu, Xiangyu Zhang, and Kouji Kozaki. Report on the First Knowledge Graph Reasoning Challenge 2018. In Xin Wang, Francesca Alessandra Lisi, Guohui Xiao, and Elena Botoeva, editors, **Semantic Technology**, Lecture Notes in Computer Science, pp. 18–34, Cham, 2020. Springer International Publishing.

[7] Takahiro Kawamura, Shusaku Egami, Kyoumoto Matsushita, Takanori Ugai, Ken Fukuda, and Kouji Kozaki. Contextualized Scene Knowledge Graphs for XAI Benchmarking. In **Proceedings of the 11th International Joint Conference on Knowledge Graphs**, IJCKG '22, pp. 64–72, New York, NY, USA, February 2023. Association for Computing Machinery.

[8] Kouji Kozaki, Shusaku Egami, Kyoumoto Matsushita, Takanori Ugai, Takahiro Kawamura, and Ken Fukuda. Datasets of mystery stories for knowledge graph reasoning challenge. In **Proc. Joint Workshops Tutorials 20th Eur. Semantic Web Conf.(ESWC)**, Vol. 3443, pp. 1–15, 2023.

[9] Christian Chiarcos. Towards interoperable discourse annotation. Discourse features in the Ontologies of Linguistic Annotation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)**, pp. 4569–4577, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA).

[10] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The Penn Discourse TreeBank 2.0. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, and Daniel Tapias, editors, **Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)**, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[11] Rasha Ali and Ban Sharief Mustafa. A Survey on Event Extraction from Webpage. In **2022 8th International Conference on Contemporary Information Technology and Mathematics (ICCITM)**, pp. 159–164, August 2022.

[12] Wei Xiang and Bang Wang. A survey of event extraction from text. **IEEE Access**, Vol. 7, pp. 173111–173137, 2019.

[13] Brenda Santana, Ricardo Campos, Evelin Amorim, Alípio Jorge, Purificação Silvano, and Sérgio Nunes. A survey on narrative extraction from textual data. **Artificial Intelligence Review**, Vol. 56, No. 8, pp. 8393–8435, August 2023.

[14] Felix Hamborg, Corinna Breitinger, and Bela Gipp. Giveme5w1h: A universal system for extracting main events from news articles. In **7th International Workshop on News Recommendation and Analytics**, pp. 35–43, 2019.

[15] Wei Xiang and Bang Wang. A Survey of Implicit Discourse Relation Recognition. **ACM Computing Surveys**, Vol. 55, No. 12, pp. 258:1–258:34, March 2023.

[16] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. **Advances in neural information processing systems**, Vol. 35, pp. 22199–22213, 2022.

[17] Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. Ragas: Automated evaluation of retrieval augmented generation, 2023.

[18] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. Www'18 open challenge: Financial opinion mining and question answering. In **Companion Proceedings of the The Web Conference 2018**, WWW '18, p. 1941–1942, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

# A 付録

**Listing 1** 特定のチャンクと特定の談話関係にあるチャンクを取得する SPARQL クエリのテンプレート

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-
    syntax-ns#>
PREFIX ex: <http://example.org/>
PREFIX pdtb: <http://purl.org/olia/discourse/
    discourse.PDTB.owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema
    #>
SELECT DISTINCT ?chunk ?text WHERE {
   {
       ?relation pdtb:hasArg1 ex:{chunk_id} .
       ?relation pdtb:hasArg2 ?chunk .
       ?chunk rdfs:label ?text .
       ?relation pdtb:hasSense ?sense .
       ?sense rdfs:subClassOf* ?sense2 .
       ?sense2 rdfs:label "{discourse_relation}" .
   } UNION {
       ?relation pdtb:hasArg2 ex:{chunk_id} .
       ?relation pdtb:hasArg1 ?chunk .
       ?chunk rdfs:label ?text .
       ?relation pdtb:hasSense ?sense .
       ?sense rdfs:subClassOf* ?sense2 .
       ?sense2 rdfs:label "{discourse_relation}" .
   }
}
```

**Listing 2** 5W1H をもとにチャンクを取得する SPARQL クエリのテンプレート

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-
    syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema
    #>
PREFIX ex: <http://example.org/>
SELECT ?chunk ?text WHERE {
   ?chunk rdf:type ex:Chunk .
   ?chunk rdfs:label ?text .
   # (start) Adding based on extraction results
   optional { ?chunk ex:{who||when||where||what||
       why||how||predicat} {variable} .
   filter({variable} = ex:{who||when||where||what
       ||why||how||predicat})
   # (end) Adding based on extraction results
} ORDER BY DESC(?who) DESC(?predicate) DESC(?what)
     DESC(?when) DESC(?where) DESC(?why) DESC(?how
    )
```

**Listing 3** テキストから 5W1H 情報を抽出するプロンプトのテンプレート

```
Extract the information "who,␣when,␣where,␣why,␣
    how,␣what,␣and␣predicate" from the following
    text.
* Please answer with the minimum number of words
    required.
* If the appropriate word is not in the provided
    text, answer "none".
* The output must be JSON format.

Text
{}

Answer:
```

**Listing 4** チャンク間の談話関係を認識するプロンプトのテンプレート

```
Select the discourse relationship between the
    following two texts from the list.
Output only the words in the list and no other
    description of any kind.

Text 1: {}
Text 2: {}

List of discourse relations: {}

Answer:
```

**Listing 5** 質問の応答に必要な談話関係を推定するプロンプトのテンプレート

```
What discourse relations are included in the
    appropriate knowledge to answer the question,
"{question}"
Select a discourse relationship from the list
    below.
The list of discourse relations: {vocabularies}

Answer: ’ Lets think step by step.</s>
```